Analisi di mercato attraverso tecniche di clustering

Progetto di Data Science Lab

Bruni Lorenzo, 886721
Geijo Bartoli Diego, 887208
Puccinelli Niccolò, 881395

---

22 settembre 2022

Indice

1 Introduzione

L'obiettivo di questa analisi è quello di effettuare una segmentazione di mercato, suddividendolo in gruppi di consumatori considerati omogenei. In questo modo è possibile, per un marketing manager, scegliere il segmento target al quale mirare e stabilire servizi e prodotti sulla base delle esigenze specifiche degli utenti appartenenti a tale gruppo. Un'implementazione corretta di tale segmentazione piò portare diversi vantaggi, tra cui: una migliore comprensione delle differenze tra i consumatori e una rafforzata corrispondenza tra i punti di forza di un azienda e le esigenze dei consumatori, tale da garantire un ritorno dell'investimento più elevato.

La prima fase dell'analisi di segmentazione è la definizione dell'obiettivo, che nel nostro caso riguarda il tentativo di migliorare l'efficacia delle campagne pubblicitarie. Pertanto, andando ad analizzare le caratteristiche che contraddistinguono ogni segmento, possiamo personalizzare il messaggio e le modalità di contatto (e.g. in base all'orario del giorno e al sistema operativo utilizzato) per ogni target di clienti.

L'approccio impiegato per la costruzione di questi segmenti è a posteriori, utilizzando tecniche statistiche di raggruppamento per la formazione di gruppi di consumatori (**clustering**). Tale impostazione non si affida a criteri predefiniti, ma consiste nel raggruppamento degli utenti in funzione delle loro similitudini, utilizzando informazioni sui consu-

matori riguardanti il loro rapporto con l'offerta dell'azienda. Questo approccio si articola in 5 fasi:

1. Selezione delle variabili di segmentazione. Nel nostro caso è stata utilizzata la percentuale di tempo trascorsa da ogni consumatore per ogni categoria di prodotto offerta dall'azienda.
2. Raggruppamento degli intervistati, che avviene mediante la scelta dell'algoritmo di raggruppamento, l'analisi della stabilità dei gruppi identificati e la scelta della segmentazione finale.
3. Profilazione dei segmenti, ovvero l'identificazione delle variabili che contribuiscono maggiormente alla creazione dei segmenti.
4. Descrizione dei segmenti, individuando quali siano le caratteristiche personali dei consumatori per cui questi differiscono in modo significativo. Tale fase si differenzia dalla profilazione (in cui si scelgono le variabili per produrre i segmenti) poichè si utilizzano altre variabili per contraddistinguere i consumatori che appartengono ai diversi segmenti.
5. Valutazione dell'utilità dei segmenti di mercato e formulazione di attività di marketing mirate.

2 Descrizione del dataset

Il dataset originale è composto da 1000 file *csv*, ognuno con 1372 colonne e circa 23.000 righe. Le colonne sono le seguenti:

- **1**: [*string*] ID utente.
- **2**: [*NaN*] Utenti sospetti.
- **3-5**: [*bool*] Interazioni degli utenti (Clicks, Impressions, Buy).
- **6-12**: [*bool*] Tipo di OS (Android, BSD, iOS, Linux, OSX, Windows, Other).
- **13**: [*int*] Tipo di device (Mobile = 1, Desktop = 2, Unknown = 3, Tablet = 5).
- **14-23**: [*bool*] Tipo di browser (Android, Chrome, Chromium, Edge, Firefox, Internet Explorer, Opera, Other, Safari, Unknown).
- **24-31**: [*float*] Attività durante gli orari del giorno, con distinzione tra weekend e giorni feriali.
- **32-38**: [*float*] Attività durante gli orari del giorno, senza distinzione tra weekend e giorni feriali.
- **39-47**: [*float*] Attività dell'utente in base al numero di caratteri dell'articolo.
- **48-73**: [*float*] Categorie dei cookies di primo livello (i.e. più generiche). Ogni elemento corrisponde a una percentuale di attività dell'utente. La somma di tali colonne per ogni riga è infatti pari a 100.
- **74-432**: [*float*] Categorie dei cookies di secondo livello.
- **433-1259**: [*float*] Categorie dei cookies di terzo livello.
- **1260-1263**: [*float*] Sentiment generico dell'utente.
- **1264-1372**: [*float*] Sentiment specifico dell'utente.

3 Selezione delle colonne

Al fine di rendere praticabile l'analisi in termini computazionali, è stata necessaria una selezione delle colonne, escludendo anzitutto le colonne relative al numero di caratteri, ai sentiment specifici (colonne 1264-1372) e alle categorie di secondo e terzo livello (colonne 74-1259). Inoltre:

- L'ID utente (colonna 1) non ci fornisce alcuna informazione utile, salvo poi per l'eventuale successivo rintracciamento dell'utente.
- La colonna 2, relativa agli utenti sospetti, è composta da soli valori nulli.
- Le colonne 3-5 (*Clicks*, *Impressions*, *Buy*) sono costituite, rispettivamente, da soli 0, 1 e 0, senza alcuna variazione.
- Sono state mantenute le colonne relative agli OS piuttosto che quelle concernenti il browser utilizzato.
- Le colonne *BSD* e *Other* sono composte da soli zeri.

- Sono state mantenute le colonne relative all'orario del giorno con distinzione tra weekend e giorni feriali, piuttosto che le successive (senza distinzione).
- Le colonne 1260-1263 (sentiment generico) sono per la quasi totalità pari a zero.

Pertanto, secondo questi criteri, il dataset è stato ridotto a 40 colonne, comprendenti il tipo di OS, il tipo di device, l'orario del giorno con distinzione tra giorni festivi e giorni feriali e le categorie di interesse di primo livello.

4 Data exploration

Analizziamo, grazie ai grafici seguenti (Figg. 1, 2, 3, 4), le proporzioni percentuali relative a tipo di OS, tipo di device, orario di attività e tipo di categoria di interesse.

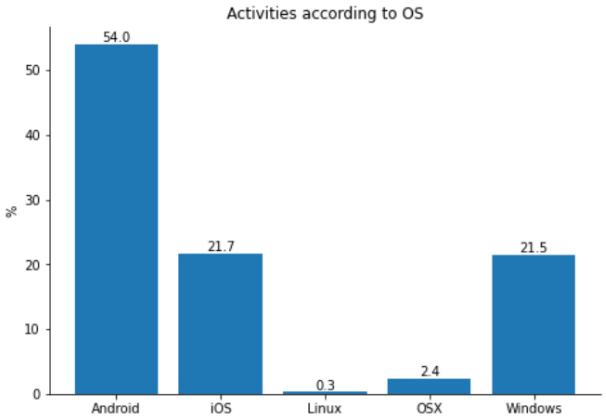

Figura 1: *OS.*

Il tipo di OS più diffuso è senza dubbio Android (54%), seguito da iOS (21.7%) e Windows (21.5%). In particolare, notiamo l'assoluta predominanza dei sistemi operativi per mobile, come confermato dal grafico successivo.

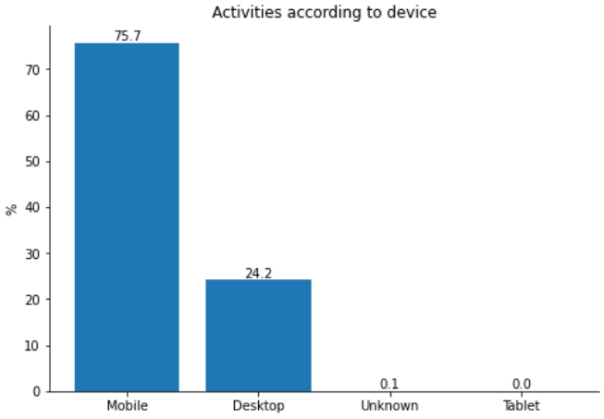

Figura 2: *Device.*

Il 75.7% ha utilizzato un dispositivo mobile, contro il 24.2% degli utenti desktop. Quasi nulle invece le categorie *Unknown* e *Tablet*.

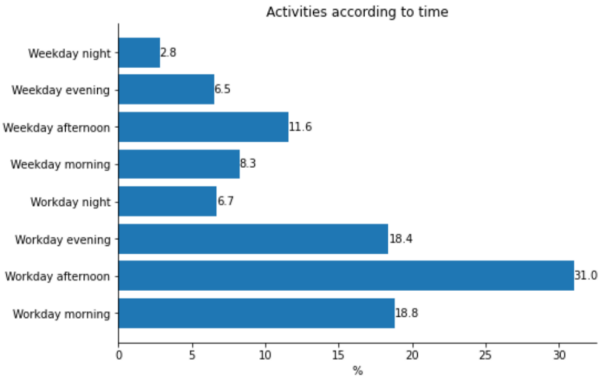

Figura 3: *Time.*

Prevedibilmente, la fascia pomeridiana è quella di maggiore attività, mentre quella con minori interazioni è la fascia notturna. Allo stesso modo, notiamo minore attività durante il weekend in generale, dovuta al fatto che vengono considerati solo gli ultimi 2 giorni della settimana invece dei restanti 5. Infine, è interessante notare la differenza tra la fascia giornaliera e la fascia serale per quanto riguarda giorni di lavoro e giorni festivi. Infatti, per quest'ultimi la differenza tra le due fasce è molto più marcata (+1.7% della fascia giornaliera), diversamente dai giorni lavorativi (differenza dello 0.4%).

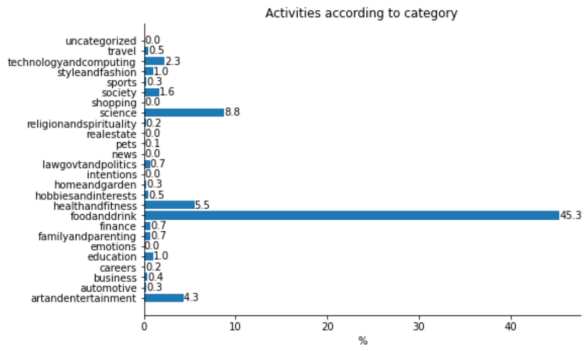

Figura 4: *Categories.*

Ben il 60.7% delle attività degli utenti riguarda la categoria relativa a cibo e bevande. Degne di nota anche le categorie *Science* (11.7%), *Healthandfitness* (7.4%) e *Artandentertainment* (5.8%). Queste 4 categorie costituiscono l'85.6% del totale. Inoltre, la categoria *Emotions* risulta nulla, pertanto è stata rimossa, insieme alla categoria *Uncategorized*, non ascrivibile ad alcun tipo di interesse.

Riportiamo ora (Figg. 5, 6) alcune osservazioni particolari sul tempo di attività e i device e gli OS utilizzati.

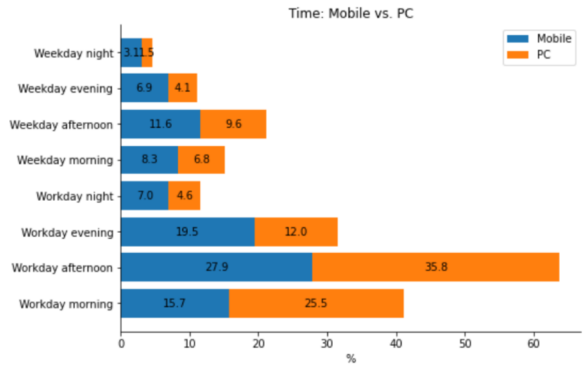

Figura 5: *Time: Mobile vs. Desktop.*

Come era lecito aspettarsi, gli utenti PC sono generalmente più attivi durante le ore di lavoro (mattina (+9.8%) e pomeriggio (+7.9%) dei giorni feriali). Gli utenti mobile sono invece molto più attivi durante le ore serali (+7.5%) e notturne (+2.4%) dei giorni lavorativi e durante tutto il weekend in generale (+7.9%).

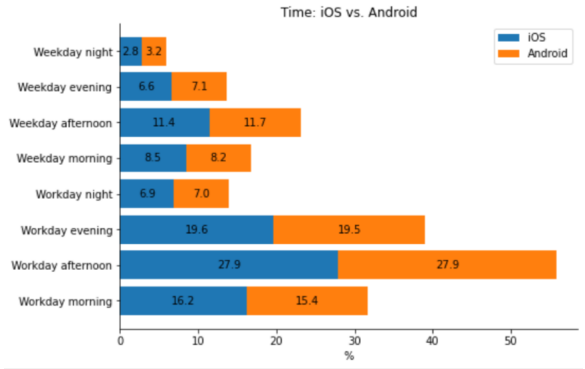

Figura 6: *Time: iOS vs. Android.*

All'interno della categoria mobile, non notiamo differenze molto sostanziali tra i due sistemi operativi. Vi sono tuttavia alcune variazioni:

- Utenti Android più attivi durante la sera e durante la notte nel weekend (10.3% vs. 9.4%).
- Utenti Android più attivi durante il weekend (30.2% vs. 29.3%).
- Utenti Android più attivi durante la notte (10.2% vs. 9.7%).
- Utenti iOS più attivi durante la mattina (24.7% vs. 23.6%).

5 Selezione delle righe

A fronte dell'esplorazione del dataset, è possibile selezionare un numero di righe minore, in modo da guadagnare tempo e spazio di computazione senza impattare eccessivamente sulla significatività dei dati. Perciò:

- Diversi utenti presentano valori nulli all'interno delle colonne relative al tempo di attività. In quanto parte centrale della nostra analisi, suddette righe sono state rimosse.
- Sono state rimosse le righe riguardanti *Tablet* e *Unknown* nella colonna relativa al tipo di device. Quest'ultime infatti costituiscono solo lo 0.1% del dataset. Per quanto riguarda il dispositivo impiegato, il confronto sarà infatti eseguito tra utenti mobile e utenti desktop.

- Come step conclusivo prima della cluster analysis, è stata effettuata un'ultima selezione. Sono infatti state eliminate le righe la cui somma delle categorie di interesse fosse inferiore al 90% (i.e. *Uncategorized* > 10%).

Grazie a questa selezione, il numero di righe è stato ridotto da circa 23 milioni a 8.586.217, sulle quali poter effettuare una più appropriata *cluster analysis*.

6 Cluster analysis

Questa tipologia di analisi consiste nella combinazione di tecniche di classificazione in grado di scomporre un insieme eterogeneo di unità statistiche in sottoinsiemi omogenei al loro interno e mutualmente esclusivi fra loro. Nell'analisi di mercato viene tipicamente impiegata per segmentare un insieme di consumatori, al fine di adottare strategie di marketing più efficaci.

La scelta dell'algoritmo di clustering e dei relativi parametri ha un impatto importante sui risultati dell'analisi. La selezione di una specifica tecnica di clustering influisce sui risultati e la segmentazione risultante da un algoritmo non può essere considerata l'unica vera soluzione di segmentazione per un determinato set di dati.

Per l'analisi di dataset di grandi dimensioni, come nel nostro caso, l'utilizzo di algoritmi di clustering gerarchici è particolarmente sconsigliato. In tali casi i dendrogrammi risultano difficili da leggere e la computazione della matrice di distanze non si adatta alla memoria del computer. Per tale motivo, per la nostra analisi è stato scelto di utilizzare una tipologia di algoritmo non gerarchico: il metodo delle **k-medie**.

6.1 PCA

Prima dell'analisi vera e propria, abbiamo ritenuto opportuna una riduzione della dimensionalità, dopo aver estratto le componenti principali tramite Principal Component Analysis. In particolare, sono state scelte le prime 6 componenti, che spiegano circa il 90% della variabilità. Tuttavia, al fine di visualizzare i cluster tramite *scatter plot*, è stata effettuata un'ulteriore analisi, considerando solo le prime due componenti, che spiegano circa il 73% della variabilità.

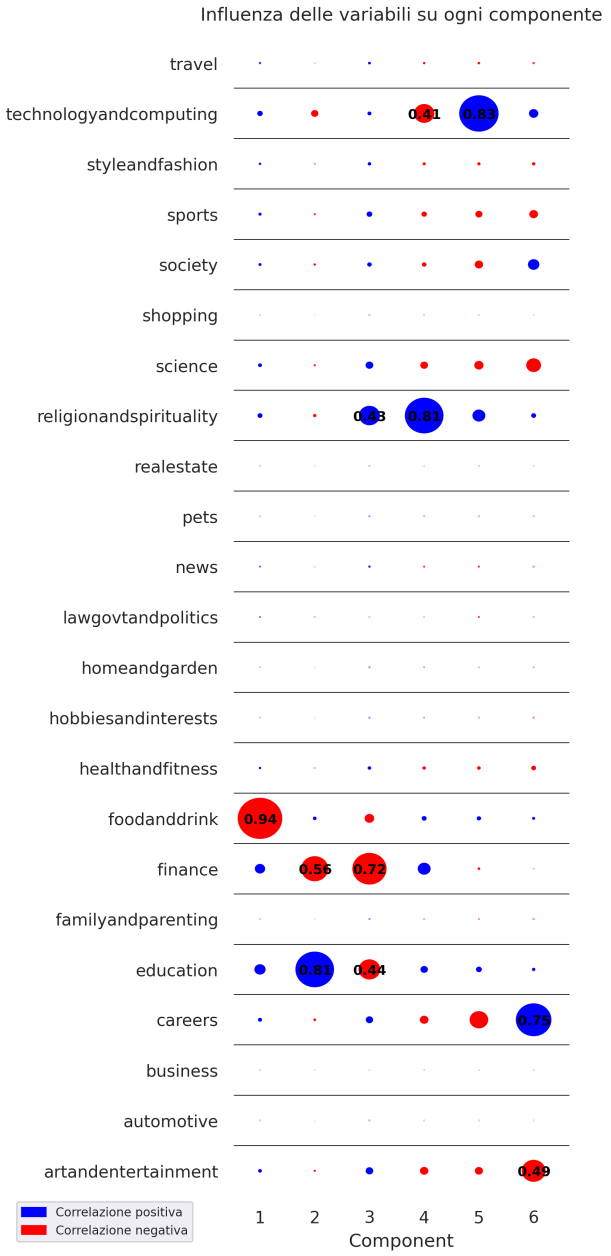

Figura 7: *Influenza delle variabili originali sulle prime 6 componenti.*

Grazie alla figura 7 è possibile visualizzare graficamente le correlazioni tra le prime 6 componenti considerate e le rispettive variabili che le compongono. Questo ci permetterà inoltre di spiegare i risultati dell'algoritmo di clustering con un maggiore grado di interpretabilità.

Le variabili con valore assoluto di correlazione più alto (≥ 0.4) per ogni componente sono:

- Componente 1: *foodanddrink* (-0.94).
- Componente 2: *education* (0.81), *finance* (-0.56).
- Componente 3: *finance* (-0.72), *education* (-0.44), *religionandspirituality* (0.43).
- Componente 4: *religionandspirituality* (0.81), *technologyandcomputing* (-0.41).
- Componente 5: *technologyandcomputing* (0.83).
- Componente 6: *careers* (0.81), *artandentertainment* (-0.56).

6.2 K-Medie

L'algoritmo delle k-medie è uno dei principali algoritmi di clustering non gerarchici e consente di raggiungere velocemente la soluzione ottimale. Poichè per l'inizializzazione di tale algoritmo è necessario specificare a priori il numero di gruppi da formare, abbiamo scelto l'**Elbow Method** come tecnica per individuare il numero ottimale di cluster. Abbiamo dunque applicato le k-medie sia nel caso delle 2 componenti, sia nel caso delle 6 componenti, comparando successivamente i risultati.

6.2.1 The Elbow Method

Per determinare il numero ottimale di cluster, decidiamo di variarne il numero (K) da 1 a 8 e, per ogni valore, calcoliamo il WCSS (*Within-Cluster Sum of Square*). Il WCSS è la somma al quadrato degli scarti di ogni osservazione con il centroide del suo cluster. Una volta rappresentato il numero di cluster con il WCSS, la curva risultante avrà la forma di un gomito e il valore del numero di cluster ottimale è rappresentato dal punto in cui cambia il flesso della curva. Da tale punto in poi l'aumento del numero di cluster non comporterà una riduzione significativa del WCSS.

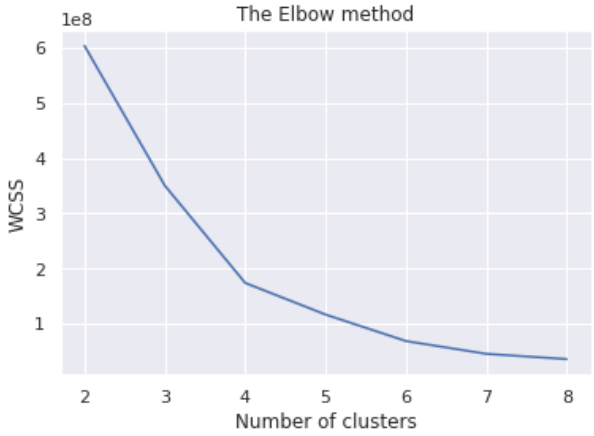

Figura 8: *The Elbow Method (2 componenti).*

Dalla figura 8 vediamo come il numero di cluster ottimale risultante sia 4. Tuttavia, successivamente, abbiamo deciso di condurre un'analisi a parte con un numero di cluster minore (3), poichè questi non risultavano ben distinti in fase di descrizione per le variabili di contatto a disposizione.

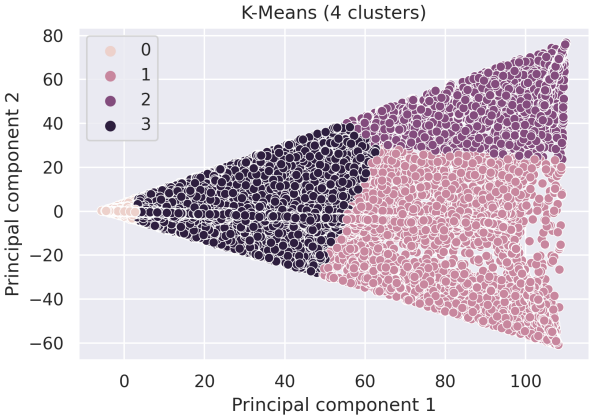

Figura 9: *K-Means (4 cluster, 2 componenti).*

La figura 9 mostra i cluster ottenuti in seguito all'addestramento dell'algoritmo delle k-medie, esplicitando 4 come numero di cluster. In tale caso possiamo notare come i cluster 0 e 3 non siano così ben diversi fra loro ed è facile supporre che, nel caso in cui scegliessimo di adottare un numero di cluster inferiore, questi 2 cluster possano fondersi.

I cluster 0, 1, 2 e 3 racchiudono, rispettivamente, le seguenti percentuali di osservazioni totali: 13.01%, 1.39%, 0.43% e 85.17%. Lo sbilanciamento dei cluster è giustificato dalla forte sparsità della matrice e dall'impiego delle k-medie, il cui obiettivo non è la produzione di cluster bilanciati. Inoltre, il dataset stesso presenta un divario importante tra *foodanddrink* e gli altri interessi, che, come vedremo, contribuisce a raggruppare gli utenti di tale categoria (ovvero una grande parte della popolazione) nel solito cluster. Questo permette inoltre di individuare relativamente piccoli gruppi di utenti con preferenze e abitudini piuttosto ben definiti, grazie ai quali poter formulare mirate attività di marketing.

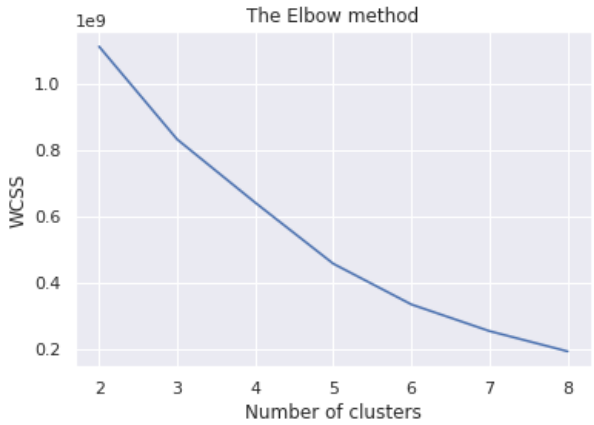

Figura 10: *The Elbow Method (6 componenti).*

Nell'ipotesi in cui scegliessimo di utilizzare 6 componenti principali come variabili per la costruzione dei cluster, notiamo come il numero di cluster ottimale sia 5 (Fig. 10). In questo caso, a differenza del precedente in cui sono state utilizzate solo 2 componenti come variabili di segmentazione, non è possibile proiettare graficamente i cluster ottenuti dall'algoritmo.

Come nel caso precedente, i cluster sono sbilanciati e contengono, rispettivamente, le seguenti percentuali di osservazioni totali: 85.63% (0), 0.94% (1), 0.43% (2), 12.53% (3), 0.47% (4).

6.2.2 Selezione di 3 cluster

In seguito, è stato applicato l'algoritmo anche per soli 3 cluster. Grazie a questo ulteriore approccio, a fronte di un inevitabile aumento del WCSS, ci aspettiamo una maggiore diversità delle osservazioni, che risulterà utile in fase di descrizione e valutazione dei segmenti. I risultati sono riportati nella sezione seguente, assieme ai valori ottenuti con la corrispettiva analisi basata sull'*elbow method*.

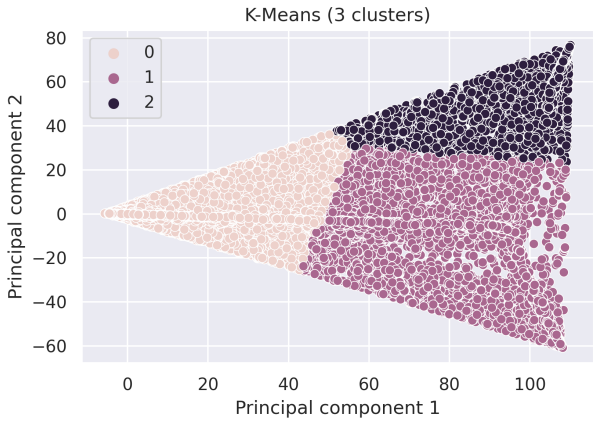

Figura 11: *K-Means (3 cluster, 2 componenti).*

Come si può notare dalla figura 11, la diminuzione del numero di cluster ha pressoché portato alla fusione dei gruppi 0 e 3, che ora raccolgono il 98.2% delle osservazioni, mentre i cluster 1 e 2 sono formati da, rispettivamente, lo 0.44% e l'1.44% degli utenti. Questi due esigui gruppi, comunque composti da, rispettivamente, 37.779 e 123.641 osservazioni, rappresentano una fetta di popolazione potenzialmente rilevante ai fini dell'analisi di marketing.

6.3 Valutazione dei risultati

Una volta raggruppati gli utenti tramite clustering, abbiamo valutato i risultati, analizzandoli da un punto di vista interno (profilazione dei segmenti, i.e. composizione dei cluster) ed esterno (descrizione dei segmenti, i.e. individuazione delle differenze mediante altre variabili, quali OS e device utilizzati).

6.3.1 Profilazione dei segmenti

Allo scopo di individuare le variabili originali che compongono i vari gruppi, abbiamo calcolato le coordinate dei centroidi di ogni cluster, per poi confrontarle con le correlazioni tra le categorie e le 6 componenti (Fig. 7).

Analizziamo anzitutto il caso a 2 componenti (Tab. 1).

	Cluster			
Coord.	**0**	**1**	**2**	**3**
x	-3.60	95.79	107.81	10.09
y	0.09	-22.67	71.12	-0.57
x	-1.86	94.19	106.87	-
y	0.004	-21.98	70.59	-

Tabella 1: *Coordinate dei centroidi calcolate tramite K-Means (2 componenti).*

Il cluster 0, comprendente la maggior parte delle osservazioni, è quello più vicino all'origine per ogni componente, risultando dunque affine al dataset di partenza. I cluster 1 e 2 suggeriscono invece una massiccia presenza di utenti negativamente correlati con *foodanddrink*, mentre differiscono per quanto riguarda le categorie *education* e *finance*, correlate alla seconda componente con coefficienti rispettivamente 0.81 e -0.56. Perciò, i cluster 1 e 2 godono di un certo livello di complementarietà per ciò che concerne *education* e *finance*. Il cluster 3 è invece molto più affine al cluster 0 rispetto ai due precedenti, per cui la loro fusione sembra garantire una migliore suddivisione.

Esaminiamo ora la tabella 2, contenente le coordinate dei centroidi con 6 componenti.

	Cluster				
Coord.	**0**	**1**	**2**	**3**	**4**
x	-3.58	90.42	107.94	104.77	10.16
y	0.09	-7.76	71.28	-51.32	-0.58
z	-0.35	28.76	-24.72	-39.67	2.60
u	0.21	-2.34	4.76	6.57	-1.65
v	0.12	-1.43	2.62	4.08	-0.96
w	0.18	3.03	1.00	0.26	-1.48
x	-1.86	94.00	106.95	-	-
y	0.004	-21.92	70.53	-	-
z	0.02	5.99	-24.41	-	-
u	-0.03	0.47	4.68	-	-
v	0.02	0.37	2.54	-	-
w	-0.03	1.89	1.07	-	-

Tabella 2: *Coordinate dei centroidi calcolate tramite K-Means (6 componenti).*

Come per il clustering a 2 componenti, il cluster 0 rappresenta maggiormente il dataset di partenza, salvo un leggero picco negativo per quanto riguarda la prima componente. Il cluster 1 mostra invece un valore superiore a 90 per x e suggerisce la presenza di utenti affini alla categoria *religionandspirituality* (correlazione di 0.43 con z), soprattutto per quanto riguarda il clustering a 5 gruppi. Il gruppo 2 assume valori simili tra i due approcci, con valori positivi notevoli per x e y e valore negativo per z. Il cluster 3 indica la presenza di utenti assimilabili alle categorie correlate positivamente con x e (parzialmente) con u e negativamente con y e z. Infine, anche per l'opzione a 6 componenti, l'ultimo cluster (4) appare assimilabile al primo.

6.3.2 Descrizione dei segmenti

Tramite l'ausilio di tabelle (Appendice A) e grafici, vediamo ora come si differenziano gli utenti tra i vari cluster dal punto di vista esterno. Analizziamo dunque la distribuzione delle osservazioni sulla base di 3 variabili esterne, riguardanti il device impiegato (mobile, desktop), la tipologia di OS installato (Android, iOS, Linux, OSX, Windows) e l'orario del gior-

no (con la distinzione tra giorni feriali e giorni festivi).

Device

Iniziamo con le figure e le tabelle che fanno riferimento all'analisi riguardante la tipologia di device utilizzato (mobile o desktop).

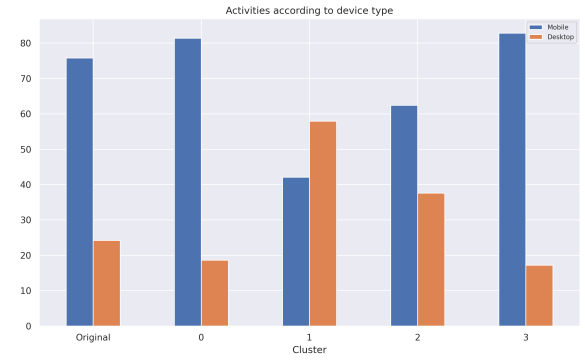

Figura 12: *Device - confronto (%) tra i cluster ottenuti tramite K-Means (4 cluster, 2 componenti).*

In figura 12, ovvero il caso con 4 cluster e 2 componenti considerate, osserviamo come i cluster 0 e 3 risultino molto simili tra loro. Il cluster 1 è invece l'unico che presenta un maggior numero di valori *Desktop* (57.93%) che di valori *Mobile* (42.97%), mentre il gruppo 2 presenta un rapporto a favore dei dispositivi mobile (62.45%), seppure più equilibrato rispetto al dataset originale.

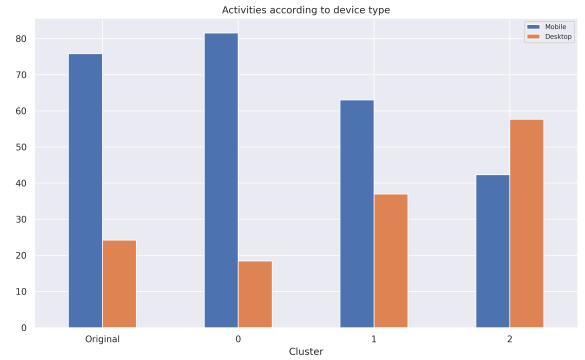

Figura 13: *Device - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 2 componenti).*

Nel caso con 3 cluster, sempre con 2 componenti (Fig. 13), notiamo che il cluster 0 è quello che presenta una maggior differenza nel numero di osservazioni tra i due valori, a favore dei valori *Mobile* (81.57%). Di nuovo, un solo cluster presenta un maggior numero di osservazioni con valore *Desktop*, il cluster 2 (57.81%). La riduzione del numero di cluster ci ha permesso di eliminare dall'analisi la somiglianza che precedentemente sussisteva tra gruppo 0 e gruppo 3.

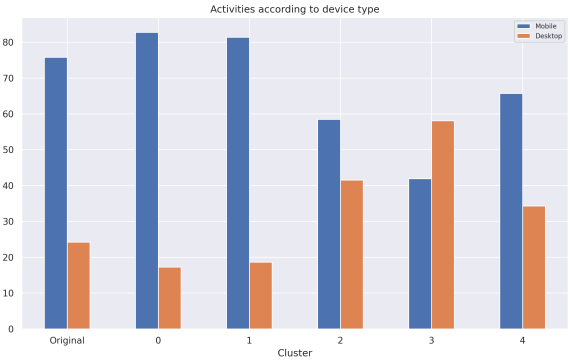

Figura 14: *Device - confronto (%) tra i cluster ottenuti tramite K-Means (5 cluster, 6 componenti).*

La figura 14 fa riferimento al caso con 5 cluster, ottenuti tramite *elbow method* considerando le prime 6 componenti. Osserviamo che il cluster 0 e il cluster 1 sono molto simili tra loro, mentre, di nuovo, un solo cluster (il 3) presenta un maggior numero di osservazioni con valore *Desktop* (58.08%).

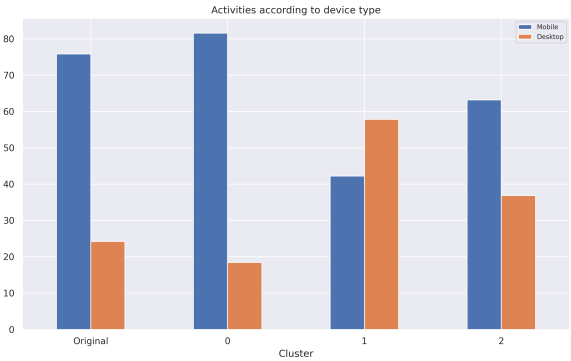

Figura 15: *Device - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 6 componenti).*

In figura 15 si passa nuovamente a 3 cluster, ma ottenuti considerando le prime 6 componenti. I gruppi sono ben distinti tra loro e

troviamo ancora un cluster con un maggior numero di osservazioni (57.81%) con valore *Desktop* (gruppo 1).

In generale, osserviamo che quando il numero di cluster è maggiore di 3 si ottengono gruppi con distribuzione delle osservazioni per *Mobile* e *Desktop* molto simile, sia per il caso a 2 componenti che per il caso a 6. Pertanto, il raggruppamento a 3 cluster risulta preferibile in termini di diversità dei gruppi e utilità dei segmenti di mercato.

OS

Proseguiamo analizzando i cluster in relazione alla tipologia di OS (Android, iOS, Linux, OSX, Windows).

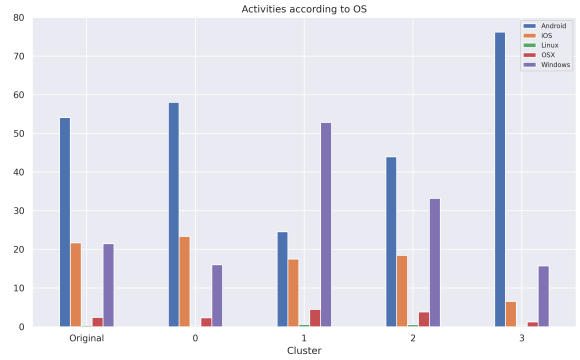

Figura 16: *OS - confronto (%) tra i cluster ottenuti tramite K-Means (4 cluster, 2 componenti).*

La figura 16 riassume il caso con 4 cluster e 2 componenti. Osserviamo che il cluster 1 è l'unico in cui l'OS con maggior numero di osservazioni non è Android (24.58%), ma Windows (52.85%), probabilmente perché la maggior parte dei device in questo cluster è di tipo *Desktop*. Il cluster 3 risulta invece molto interessante per la proporzione tra utenti Android (76.27%) e utenti iOS, pari a solo il 6.55%.

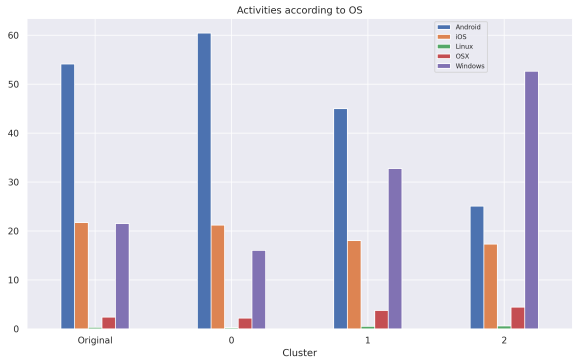

Figura 17: *OS - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 2 componenti).*

Passiamo alla figura 17, il caso con 3 cluster e 2 componenti. Non osserviamo più un cluster con un'ampia differenza di cardinalità tra Android e iOS, mentre gli altri cluster risultano molto simili al raggruppamento precedente. In questo caso, la diminuzione del numero di cluster coincide con una perdita di informazione. Il raggruppamento precedente infatti può essere utile qualora l'attività di marketing miri a dispositivi Android, a scapito di device mobili basati su iOS.

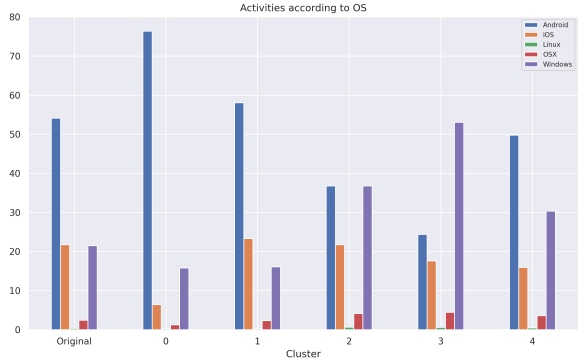

Figura 18: *OS - confronto (%) tra i cluster ottenuti tramite K-Means (5 cluster, 6 componenti).*

La figura 14 descrive il raggruppamento con 5 cluster e 6 componenti. Abbiamo nuovamente un cluster (gruppo 0) in cui il numero di osservazioni con OS Android si distacca chiaramente dal resto (76.41%), e un solo cluster (gruppo 3) che non presenta Android come OS con cardinalità maggiore, bensì Windows (53.02%). Tuttavia, l'incremento del numero di cluster ha causato una diminuzio-

ne delle differenze tra questi. I cluster 1 e 2 di Fig. 16 sembrano infatti essersi divisi nei cluster 2, 3 e 4.

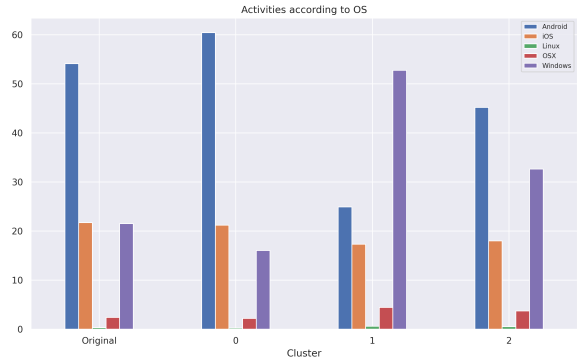

Figura 19: *OS - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 6 componenti).*

Per questa tipologia di raggruppamento (3 clusters e 6 componenti, Fig. 19), molto simile al caso con 3 cluster e 2 componenti, osserviamo che il cluster 1 è l'unico in cui l'OS con cardinalità maggiore è Windows (52.76%) e non Android (24.91%).

Nel complesso, concludiamo che, quando il numero di cluster è maggiore di 3, si rileva la presenza di un gruppo in cui la differenza di cardinalità tra Android e gli altri OS cresce notevolmente. Inoltre, è sempre presente un cluster in cui il sistema operativo più osservato è Windows e non Android, probabilmente perché vengono inseriti nello stesso gruppo molti device di tipo *Desktop*.

Time

Concludiamo con l'analisi dei raggruppamenti sulla base di giorno (feriale o festivo) e orario. Naturalmente, i valori di tipo *Workday* presentano una cardinalità maggiore rispetto ai valori *Weekday*, in quanto questi ultimi rappresentano solo i due giorni finali della settimana.

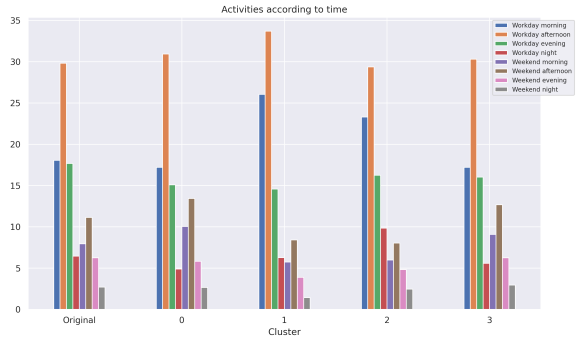

Figura 20: *Time - confronto (%) tra i cluster ottenuti tramite K-Means (4 cluster, 2 componenti).*

In figura 20 è rappresentato il caso con 4 cluster e 2 componenti. I cluster 0 e 3 presentano valori molto simili e si distinguono dagli altri per un maggiore quantitativo di utenti attivi durante la mattina (rispettivamente 10.03% e 9.07%) e il pomeriggio (rispettivamente 13.42% e 12.67%) dei giorni festivi. Il cluster 1 si distingue invece per i valori delle variabili *Workday morning* (26.05%) e *Workday afternoon* (33.71%). Infine, il gruppo 2 risulta interessante per la percentuale di utenti attivi durante la notte dei giorni lavorativi (9.84%).

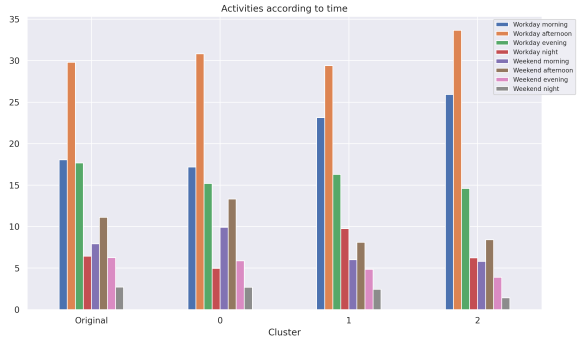

Figura 21: *Time - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 2 componenti).*

Passiamo alla figura 21, che rappresenta l'approccio con 3 cluster e 2 componenti. Notiamo che i precedenti cluster 0 e 3 probabilmente sono stati accorpati nel gruppo 0, che infatti si contraddistingue per i valori di *Weekday morning* (9.91%) e *Weekday afternoon* (13.32%). Il gruppo 1 ha la maggiore percentuale di *Workday night* (9.73%), mentre

l'ultimo cluster è quello in cui sono raggruppati gli utenti più attivi la mattina (26.01%) e il pomeriggio (33.69%) dei giorni feriali. La diminuzione del numero di cluster risulta dunque vantaggiosa, avendo eliminato la somiglianza che prima sussisteva tra i gruppi 0 e 3.

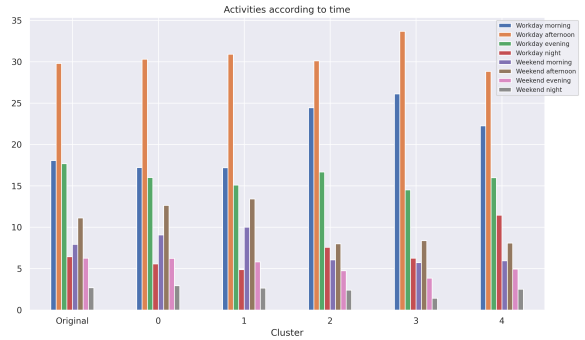

Figura 22: *Time - confronto (%) tra i cluster ottenuti tramite K-Means (5 cluster, 6 componenti).*

La figura 22 descrive il raggruppamento con 5 cluster e 6 componenti. Nei cluster 2 e 3 le osservazioni di tipo *Workday* (entrambi pari a circa l'81% del totale) si distaccano maggiormente da quelle di tipo *Weekend*, mentre nei cluster 0 e 1 accade il contrario (in cui la percentuale per il weekend è oltre il 31%). Inoltre, il gruppo 4 è interessante per il numero di osservazioni di tipo *Workday night* (10.84%). Ciò nonostante, il numero di cluster appare eccessivo, in quanto le coppie di cluster (0, 1) e (2, 3) presentano valori molto simili.

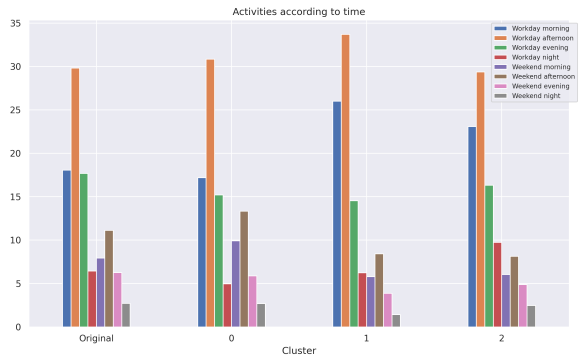

Figura 23: *Time - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 6 componenti).*

Considerando 3 cluster e 6 componenti (Fig. 23), osserviamo valori pressoché equivalenti al raggruppamento presentato in figura 21.

Sostanzialmente, per le variabili relative al tempo, l'approccio con 3 cluster risulta essere più indicato, in quanto massimizza le differenze tra i gruppi, permettendo di identificare segmenti di consumatori ben più definiti.

7 Conclusioni

Tramite questa analisi abbiamo effettuato una segmentazione di mercato al fine di garantire determinati servizi a diverse categorie di utenti, in base alle loro esigenze e abitudini specifiche. Dopo aver selezionato opportunamente le colonne e le righe per la *cluster analysis*, abbiamo adottato una riduzione della dimensionalità mediante *Principal Component Analysis*, conducendo un'analisi parallela tra le prime 2 e le prime 6 componenti. E' stato dunque applicato l'algoritmo di clustering delle k-medie, facendo ricorso all'*elbow method* per stabilire il numero ottimale di cluster (k). Tuttavia, allo scopo di evidenziare in misura maggiore le differenze tra i gruppi di utenti, abbiamo svolto un ulteriore raggruppamento con soli 3 cluster.

L'analisi proposta combina dunque 4 diversi approcci a seconda del numero di componenti considerate (2 o 6) e del numero di cluster utilizzati.

I risultati (Sez. 6.3, Appendice A) evidenziano l'identificazione di ristrette percentuali di utenti piuttosto polarizzati verso determinati interessi, orari del giorno e device e OS impiegati. In particolare, l'approccio a 3 cluster è stato in grado di individuare, mediante un minore numero di gruppi, segmenti di mercato molto diversi tra loro, sulla base dei quali è possibile formulare mirate attività di marketing.

A Appendice: Tabelle

		Cluster			
Device	Data	0	1	2	3
Mobile	75.79	81.38	42.97	62.45	82.81
Desktop	24.21	18.62	57.93	37.55	17.19
OS	Data	0	1	2	3
Android	54.09	57.98	24.58	43.98	76.27
iOS	21.71	23.41	17.52	18.48	6.55
Linux	0.32	0.23	0.60	0.52	0.21
OSX	2.39	2.32	4.45	3.81	1.26
Windows	21.49	16.06	52.85	33.21	15.71

Tabella 3: *Device e OS - confronto (%) tra i cluster ottenuti tramite K-Means (4 cluster, 2 componenti).*

		Cluster			
Workday	Data	0	1	2	3
Morning	18.06	17.19	26.05	23.29	17.19
Afternoon	29.81	30.93	33.71	29.37	33.71
Evening	17.68	15.08	14.57	16.25	16.03
Night	6.43	4.88	6.26	9.84	5.57
Tot	71.98	69.10	80.59	78.75	72.50
Weekday	Data	0	1	2	3
Morning	7.94	10.03	5.73	5.97	9.07
Afternoon	11.13	13.42	8.40	8.02	12.67
Evening	6.25	5.81	3.87	4.82	6.23
Night	2.70	2.65	1.41	2.44	2.94
Tot	28.02	31.90	19.41	21.25	17.50

Tabella 4: *Time - confronto (%) tra i cluster ottenuti tramite K-Means (4 cluster, 2 componenti).*

	Cluster				
Device	0	1	2	3	4
Mobile	82.79	81.39	59.60	41.92	64.31
Desktop	17.21	18.61	40.40	58.08	35.69
OS	0	1	2	3	4
Android	76.41	58.08	35.71	24.38	48.81
iOS	6.39	23.33	23.89	17.57	15.51
Linux	0.21	0.23	0.50	0.61	0.52
OSX	1.25	2.31	3.62	4.42	3.87
Windows	15.74	16.05	36.28	53.02	31.29

Tabella 5: *Device e OS - confronto (%) tra i cluster ottenuti tramite K-Means (5 cluster, 6 componenti).*

	Cluster				
Workday	0	1	2	3	4
Morning	17.22	17.19	25.92	26.14	21.72
Afternoon	30.31	30.92	30.83	33.71	28.60
Evening	16.01	15.09	16.33	14.51	16.28
Night	5.56	4.88	7.87	6.27	10.84
Tot	69.00	68.08	80.95	80.63	77.44
Weekday	0	1	2	3	4
Morning	9.08	10.02	5.49	5.73	6.25
Afternoon	12.65	13.42	6.92	8.38	8.66
Evening	6.23	5.82	4.40	3.85	5.09
Night	2.94	2.65	2.25	1.41	2.56
Tot	31.00	31.92	19.05	19.37	22.56

Tabella 6: *Time - confronto (%) tra i cluster ottenuti tramite K-Means (5 cluster, 6 componenti).*

		Cluster		
Device	Data	0	1	2
Mobile	75.79	81.57	63.15	42.19
Desktop	24.21	18.43	36.85	57.81
OS	Data	0	1	2
Android	54.09	60.41	45.00	25.07
iOS	21.71	21.17	18.03	17.30
Linux	0.32	0.23	0.51	0.60
OSX	2.39	2.18	3.73	4.42
Windows	21.49	16.01	32.73	52.61

Tabella 7: *Device e OS - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 2 componenti).*

	Cluster		
Device	0	1	2
Mobile	81.57	42.19	63.15
Desktop	18.43	57.81	36.85
OS	0	1	2
Android	60.41	24.91	45.18
iOS	21.17	17.31	17.98
Linux	0.23	0.60	0.51
OSX	2.18	4.42	3.72
Windows	16.01	52.76	32.61

Tabella 9: *Device e OS - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 6 componenti).*

		Cluster		
Workday	Data	0	1	2
Morning	18.06	17.19	23.08	26.01
Afternoon	29.81	30.84	29.38	33.69
Evening	17.68	15.21	16.31	14.54
Night	6.43	4.97	9.73	6.24
Tot	71.98	68.21	78.50	80.48
Weekday	Data	0	1	2
Morning	7.94	9.91	6.04	5.80
Afternoon	11.13	13.32	8.13	8.42
Evening	6.25	5.87	4.87	3.88
Night	2.70	2.69	2.46	1.42
Tot	28.02	31.79	21.50	19.52

Tabella 8: *Time - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 2 componenti).*

	Cluster		
Workday	0	1	2
Morning	17.19	26.00	23.08
Afternoon	30.84	33.69	29.38
Evening	15.21	14.54	16.31
Night	4.97	6.24	9.73
Tot	68.21	80.47	78.50
Weekday	0	1	2
Morning	9.91	5.80	6.04
Afternoon	13.32	8.42	8.13
Evening	5.87	3.88	4.87
Night	2.69	1.42	2.46
Tot	31.79	19.53	21.50

Tabella 10: *Time - confronto (%) tra i cluster ottenuti tramite K-Means (3 cluster, 6 componenti).*